STK 1000 høsten 2013
Oblig II

Dette er oppgavesettet for andre obligatoriske innleveringsprosjekt for kurset STK 1000. Det legges ut på kurssiden **tirsdag 22/10/13**, og leveringsfristen er **torsdag 31/10/13**. Besvarelsen skal leveres til instituttkontoret ved Matematisk institutt, senest kl. 14:30 den dagen. Sjekk de praktiske detaljene samlet i `www.mn.uio.no/math/studier/admin/obligatorisk-innlevering/index.html`, der det også forklares at du skal bruke en bestemt standardisert «Forside for obligatoriske innleveringer».

Hvis flere samarbeider om å løse oppgavene, må likevel hver student levere sin egen besvarelse; spesielt må hver student levere resultatene av egne simuleringer i oppgave 1. Det må gå frem av besvarelsen hvem du eventuelt har samarbeidet med.

Du kan levere håndskrevet eller maskinskrevet (tekstbehandlet) besvarelse. Der du bruker MINITAB må utskrifter og plott legges ved (eventuelt opereres inn i tekstbehandlingsdokumentet). Du vil kunne få bruk for introduksjonsheftet *Starthjelp i* MINITAB*: Versjon 16 for Windows*, som kan finnes og printes ut fra kurssiden for STK1000 fra høstsemesteret 2012, sammen med praktiske opplysninger om hvordan man kan koble seg til UiOs maskiner fra egen pc, etc.

Nils Lid Hjort

Oppgave 1

HVA ER KONFIDENS? Tenk deg at du har 20 tilfeldige utvalg, hver av størrelse 30, fra en $\mathrm{N}(8,3)$-fordelt populasjon (altså en normalfordeling med forventning 8 og standardavvik 3). For hvert av de 20 utvalgene kan du finne et 90% konfidensintervall for forventningsverdien (som vi her vet er lik 8). For enkelhets skyld antar vi at disse konfidensintervallene blir laget ut fra presist kjennskap til at standardavviket er $\sigma = 3$, slik at metodikk og formler fra lærebokens Chapter 6.1 kan anvendes direkte (se f.eks. side 345); dersom σ er ukjent og må estimeres fra det enkelte datasett trenger man en bestemt modifikasjon som tas opp i bokens Chapter 7.1 (se f.eks. side 402).

(a) Hvor mange av de 20 konfidensintervallene venter du vil inneholde den riktige verdien, altså $\mu = 8$?

(b) La Y være en stokastisk variabel som angir hvor mange av de 20 konfidensintervallene som vil inneholde den riktige verdien 8. Hvilken fordeling har Y? Begrunn ditt svar.

(c) Hva er forventning og standardavvik for Y?

– Du kan bruke MINITAB til å simulere den situasjonen som er beskrevet innledningsvis i oppgaven. Først trekker du 20×30 stokastiske variable som er uavhengige og $\mathrm{N}(8,3)$-fordelte. Dette gjør du ved å gå til `Calc`, så til `Random Data`, og så til `Normal`. Generer her 30 rader med data og ta vare på dem i kolonnene `C1-C20`. Tallene i en kolonne gir deg nå de 30 observasjonene i et tilfeldig utvalg, mens de 20 kolonnene gir deg de 20 utvalgene.

(d) Nå skal du finne et konfidensintervall for hvert av de 20 utvalgene. Kommandoen er fra `Stat` til `Basic Statistics` og `1-Sample Z`. Husk at konfidensgraden skal være 90%. Utfør kommandoene ovenfor og angi de 20 konfidensintervallene du får. Hvor mange av dem inneholder den korrekte parameterverdien? Kommenter resultatet i lys av punktene (a) og (b).

Oppgave 2

Også skøyteløpere blir slitne, når de bare har løpt langt nok. Likevel har mange skøyteløpere høyere gjennomsnittsfart på mellomdistansen 1500 meter (tre og tre-kvart runde) enn på sprintdistansen 500 meter (en og en-kvart runde). Verdensmester Håvard Bøkko, for eksempel, har 1.42.67 på 1500 m (Salt Lake City 2009), som svarer til en gjennomsnittlig 500 m-fart på 102.67/3 = 34.223, betydelig raskere enn hans beste 500 m-tid 35.85 (Hamar, sist uke). Mange andre løpere har dog høyere fart på en 500 m enn på en 1500 m. For å gripe litt bakover i tiden kan man ihukomme Knut «Kuppern» Johannesen, mannen hvis verdensrekordtid på 10.000 m er nevnt i minst syv norske skjønnlitterære romaner, og som får en statue av seg avduket ved Bislett onsdag 6. november. Hans 2.09.4 (fra Karuizawa 1963) er en del langsommere (43.133 pr. 500 m) enn hans 42.3 (fra Squaw Valley 1960).

Vi skal i denne oppgaven undersøke hvor vanlig det er at løpere har høyere tempo på 1500 m enn på 500 m, og skal også finne ut om det er en signifikant forskjell, eller om forskjellene i tempo kan regnes som tilfeldige variasjoner. Til dette formål konsulterer jeg den såkalte Adelskalenderen, over verdens beste skøyteløpere gjennom tidene, basert på deres personlige rekorder og deres sammenlagtpoengsum (som er $x_1+x_2/3+x_3/10+x_4/20$, der x_1, x_2, x_3, x_4 er tidene på de fire klassiske distanser 500 m, 1500 m, 5000 m, 10000 m); se siste side i dette oppgavesettet. For dagens formål plukker jeg ut de 33 beste (der sampelstørrelsen er valgt av skandinaviskromantiske grunner, idet jeg da får med meg ikke bare løpere fra USA, Nederland, Norge, Russland, Italia, Canada, Frankrike, Belgia, Kazakstan, Tyskland, Sør-Korea, New Zealand, Latvia, men også fra Sverige). For disse 33 helter beregner jeg brøkene

$$R = \frac{(x_2/3)}{x_1},$$

altså 500-fart under 1500-løpet, delt på 500-tiden. Observasjonene $R_1, \ldots, R_{33}$ er som følger:

```
0.968  0.954  0.959  0.955  0.956  0.972  0.958  0.972  0.955  0.956
0.961  0.975  0.953  0.961  0.952  0.976  0.983  0.952  0.965  0.962
0.966  0.957  0.923  0.974  0.953  0.953  0.963  0.958  0.949  0.999
0.967  0.973  0.952
```

(a) Lag et histogram for de 33 ratio-variablene. Beregn også gjennomsnitt, median og standardavvik. Hvem av disse 33 løperne, om noen, er potensielle outliere, i henhold til den såkalte $1.5 \times$ IQR-regelen? Kommenter det du finner ut.

(b) La μ betegne den underliggende middelverdi for hele R-populasjonen, altså R-observasjonenes teoretiske forventningsverdi. Finn et konfidensintervall for μ med konfidensgrad 99%, idet du (for enkelhets skyld) antar at R-ene kommer fra en populasjon med et standardavvik på 0.014. Gjør rede for hvilke antagelser du gjør for å kunne tolke intervallet som et 99% konfidensintervall.

(c) Vi er altså interessert i om R-ratioenes gjennomsnitt i populasjonen av gode skøyteløpere er lik 1, eller ikke. Sett opp en test for hypotesen $H_0\colon \mu = 1$, mot alternativet $H_a\colon \mu \neq 1$, idet du fortsatt antar at standardavviket for enkeltratioene er $\sigma = 0.014$. Utfør testen, og beregn (eller anslå) dessuten P-verdien. Formuler en konklusjon.

(d) Hva kan det komme av, at mange skøyteløpere altså har høyere gjennomsnittstempo på tre og tre-kvart runde, enn på en og en-kvart runde?

Oppgave 3

Det er mulig at utvalget av skøyteløpere gjort over ikke er helt representativt, når det gjelder problemstillingen om 1500 m løpes i høyere tempo enn 500 m. For å undersøke dette nærmere plukker jeg i tillegg ut 33 noe mindre glamorøse løpere, nemlig de som i øyeblikket befinner seg på plassene fra 268 til 300 på Adelskalenderen. Så beregner jeg R-brøken også for disse, og får

```
0.986  0.986  0.976  0.982  0.991  0.996  0.981  0.994  0.972  0.977
0.958  1.017  0.930  0.972  0.995  1.007  0.961  0.953  0.967  0.965
0.960  0.983  1.002  0.994  1.012  0.972  0.983  0.972  0.990  0.941
0.988  0.986  0.983
```

Disse ses igjen på som uavhengige observasjoner med et underliggende populasjonsmiddeltall μ', og for å holde oss til stoffet i lærebokens Chapter 6, uten å gripe inn i Chapter 7, antar vi at sannsynlighetsfordelingen for disse R-målingene, for disse mindre supre løpere, kan anses å ha et kjent standardavvik $\sigma' = 0.018$.

Det er naturligvis store forskjeller mellom nivået til den supergruppen jeg samplet fra først (de 33 beste i verden) og til den noe mer pedestriøse gruppen jeg studerer i denne oppgaven (fra nr. 268 til nr. 300 på Adelskalenderen). To naturlige problemstillinger er som følger. Gjelder det samme funn som for Oppgave 2, med hensyn til hypotesen om at μ' er lik 1? Og er det forskjell med hensyn til ratioen mellom 1500-fart og 500-fart?

(a) Lag boksplott for hver av de to grupper av R-brøker, gjerne i samme plott. Kommenter det du ser. Test, som i forrige oppgave, om $\mu' = 1$ eller ikke, og formuler en konklusjon.

(b) Test hypotesen at de to gruppene av skøyteløpere er like, med hensyn til ratioen mellom 1500-fart og 500-fart. Gjør rede for de antagelser du bruker, og formuler en konklusjon.

Men no skrid Kuppern i Squaw Valley!
Eg har ikkje tenkt å gå nokon 10000 m,
men ordi fær ein djervare sving,
og ho mor grip fastare um staven.

Bonusspor: Toppen av Adelskalenderen, pr. 22. oktober 2013

1	DAVIS Shani	USA	34.78	1.41.04	6.10.49	13.05.94	144.806
2	KRAMER Sven	NED	36.17	1.43.54	6.03.32	12.41.69	145.099
3	HEDRICK Chad	USA	35.52	1.42.14	6.09.68	12.55.11	145.289
4	BØKKO Håvard	NOR	35.85	1.42.67	6.09.94	12.53.89	145.761
5	SKOBREV Ivan	RUS	35.90	1.42.94	6.10.58	12.58.36	146.189
6	BLOKHUIJSEN Jan	NED	35.59	1.43.78	6.11.97	13.00.04	146.382
7	FABRIS Enrico	ITA	35.99	1.43.48	6.06.06	13.10.60	146.619
8	VERWEIJ Koen	NED	35.64	1.43.92	6.12.20	13.08.97	146.948
9	MARSICANO Trevor	USA	35.72	1.42.31	6.16.55	13.21.06	147.531
10	KUCK Jonathan	USA	35.97	1.43.12	6.16.28	13.11.24	147.533
11	UYTDEHAAGE Jochem	NED	36.27	1.44.57	6.14.66	12.58.92	147.538
12	HANSEN Brian	USA	35.33	1.43.35	6.19.14	13.24.11	147.899
13	VERHEIJEN Carl	NED	36.99	1.45.78	6.08.98	12.55.30	147.913
14	OLDE HEUVEL Wouter	NED	36.17	1.44.24	6.16.26	13.14.79	148.281
15	ERVIK Eskil	NOR	37.03	1.45.73	6.10.65	12.59.69	148.322
16	MORRISON Denny	CAN	34.85	1.42.01	6.24.13	13.41.65	148.348
17	WENNEMARS Erben	NED	34.68	1.42.32	6.28.42	13.35.67	148.411
18	CONTIN Alexis	FRA	36.66	1.44.73	6.15.74	13.12.11	148.749
19	MAKOWSKY Lucas	CAN	35.76	1.43.53	6.24.57	13.21.93	148.823
20	PEDERSEN Sverre Lunde	NOR	36.56	1.45.48	6.17.78	13.08.82	148.939
21	PARRA Derek	USA	35.88	1.43.95	6.17.98	13.33.44	149.000
22	SWINGS Bart	BEL	36.73	1.45.50	6.18.56	13.08.08	149.156
23	DANKERS Arne	CAN	37.58	1.44.01	6.14.01	13.10.58	149.180
24	TUITERT Mark	NED	35.20	1.42.87	6.27.63	13.38.91	149.198
25	de JONG Bob	NED	37.86	1.48.22	6.08.76	12.48.20	149.219
26	BABENKO Dmitrij	KAZ	36.83	1.45.26	6.20.32	13.10.31	149.463
27	LEHMANN Robert	GER	36.34	1.45.03	6.20.16	13.22.33	149.482
28	LEE Seung-Hoon	KOR	37.36	1.47.38	6.14.67	12.57.27	149.483
29	DOBBIN Shane	NZL	37.31	1.46.19	6.15.69	13.04.36	149.493
30	SILOVS Haralds	LAT	35.32	1.45.90	6.17.13	13.43.66	149.516
31	SANFRATELLO Ippolito	ITA	36.59	1.46.13	6.16.27	13.18.98	149.542
32	ROMME Gianni	NED	36.97	1.47.88	6.14.70	13.03.40	149.570
33	RÖJLER Johan	SWE	37.05	1.45.85	6.18.35	13.08.42	149.589

Tyve av disse tre-og-tredve har vunnet olympiske gullmedaljer eller verdensmesterskap.